# Dokumentation - Alignment Faking

Blog von JS und RR im Seminar "Möglichkeiten der kollaborativen Texterstellung in digitalen Zeiten".

# 1. Ablauf

1. **Themenwahl**: Liste an Themenvorschlägen (JS), aus denen RR einen wählte
2. **Verallgemeinerung des Themas**: Überlegungen zum Hauptthema, Unterthemen und zugehörigen Aspekten. Wahl konkreter Unterthemen für Posts.
3. **Zielgruppenwahl**: Überlegung, für wen das Thema verständlich und relevant sein könnte (s. nächsten Absatz)
4. **Schreiben der Texte**: Individuelle Recherche, Bildersuche, Formulierung in Markdown-Editoren
5. **Software-Installation und Technik**: Filezilla, Anmeldung in Bludit, Anlegen der Seite
6. **Upload der Texte und Bilder nach Bludit**
7. **Layout und Darstellung**: Format, Texte "Über", Titel und Untertitel der Seite etc.
8. **Angleichen der Überschriften**: Überschriften und Unterüberschriften in ähnlichen Stilen wählen
9. **Dokumentation**: Erstellen und Einfügen dieser Dokumentation

# 2. Zielgruppe

- Vor allem Studierende
- sollten fließend Deutsch und Englisch lesen können
- Vage naturwissenschaftlicher Hintergrund
- Interesse an philosophischen Diskussionen

Diese Zielgruppe schien uns geeignet, da unsere Texte sowohl auf Deutsch als auch auf Englisch geschrieben sind und teilweise auch in deutschen Texten englische Fachbegriffe nutzen ("Social Deduction Games"). Ein naturwissenschaftlicher Hintergrund zahlt sich in dem Post über Alignment Faking in LLMs ("I'll be back, says the AI") aus, da das Vorgehen beim Trainieren einer KI oder grundlegende Hintergründe nicht expliziert erklärt werden. Die Hürde sollte allerdings nicht sehr hoch sein, und ein bisschen technisches Grundwissen sollte zum Verständnis genügen.

Ein Interesse an philosophischen Diskussionen ist insofern von Vorteil, dass "Täuschung" das Thema ist, das "Alignment Faking" zu Grunde liegt, was natürlich ein sehr menschliches, aber auch ein sehr moralisch fragwürdiges Thema ist.

# 3. Thema

Das Thema ist aus dem Interesse JS' an dem entsprechenden Paper von Anthropic entstanden (siehe Post "I'll be back, says the AI"). Von dort aus haben wir das Thema in den menschlich-sozialen Bereich erweitert, also quasi als "Täuschung" verallgemeinert. Spezifisch als Täuschung, in der eine falsche Loyalität vorgetäuscht wird.

Dazu haben wir dann neben dem Alignment Faking in LLMs noch drei andere Themen oder Beispiele herausgesucht, in denen diese Art der Täuschung vorkommt: Literatur, professionelles Umfeld (Industiespionage, Phishing etc.; siehe auch "Julianus or just Ianus?") und in der heutigen Spielkultur (Social Deduction Games, siehe "Zwei Werwölfe im Dorfbewohner-Pelz").

Das Ganze haben wir einem einführenden Artikel zusammengefasst, der quasi eine Klammer über alle Unterthemen darstellt und den Zusammenhang aufzeigt ("Othello, Social Engineering, oder die Werwölfe von Venedig").

# 4. Organisation, Aufteilung und Realisierung

- **Themenvorschläge**: JS
- **Themenauswahl**: JS und RR
- **Verallgemeinerung und Festlegung der Subthemen**: JS und RR
- Zusammenfassender Post **"Othello, Social Engineering, oder die Werwölfe von Venedig"**: RR
- Post zum originalen Thema, dem Paper von Anthropic (**"I'll be back, says the AI"**): JS
- Post zu Social Deduction Spielen **"Zwei Werwölfe im Dorfbewohner-Pelz"**: RR
- Post zu Julianus (**"Julianus or just Ianus?"**): JS
- **Dokumentation** (Vorlage): JS
- **Dokumentation** (ausformuliert): RR
- **Technische Einrichtung** mit FileZilla: RR
- **Upload in Bludit**: RR
- **Layout und Gestaltung**: RR

# 5. Quellen


- "I'll be back, says the AI":
  - https://terminator.fandom.com/wiki/Skynet
  - https://www.anthropic.com/research/alignment-faking
  - https://www.anthropic.com/news/disrupting-AI-espionage
  - https://arxiv.org/abs/2412.14093
  - https://d-infinity.net/posts/game-content/starfinder-t-2-aerial-autonomous-assault-vehicle
  - https://www.bbc.com/news/articles/cly7jrez2jno
  - Bildquelle: https://www.artstation.com/artwork/WX6aNQ
- "Julianus or just Ianus?"
  - https://www.amazon.de/-/en/Nick-Holmes-ebook/dp/B0C445G943
  - https://en.wikisource.org/wiki/The_Works_of_the_Emperor_Julian/The_heroic_deeds_of_Constantius
  - Bildquelle: https://www.flickr.com/photos/23416307@N04/14954464972/in/album-72157646680119601
  - Bildquelle: https://www.musee-moyenage.fr/collection/oeuvre/julien-apostat.html?utm_source=chatgpt.com
- "Othello, Social Engineering, oder die Werwölfe von Venedig":
  - Bildquelle: Toa Heftiba, Unsplash
  - Bildquelle: Sagar Paranjape, Unsplash
- "Zwei Werwölfe im Dorfbewohner-Pelz":
  - Bildquelle: Eigene Bilder
  - Inhalt: Eigene Erfahrung